ORIE 4741: Learning with Big Messy Data

Exploratory Data Analysis

Professor Udell
Operations Research and Information Engineering
Cornell

September 3, 2019
Questions from piazza

- this course vs ORIE 4740 vs CS ML vs CS ML for Data Science
- how deeply will we understand algorithms in this class?
- how much coding?
- enrollment woes: there is hope!
Why Julia?

- the two language problem
- Julia is fast (JIT-compiled)
- Julia has pleasant syntax, especially for linear algebra (MATLAB-like, but more principled)
- Julia supports efficient parallelism (including multithreading)
- The Julia ecosystem

For this class: you can use any language you'd like (which your TAs can read), but the course staff will only support Julia.
Topics to review

We will cover (most of) these in section, too:

- Linear algebra: invertible matrices, rank, norm, basic matrix identities. When is a matrix invertible?
- QR factorization
- Gradients (multivariate derivative)
- Projections
- SVD
- Union bound
- Computational complexity
Why look at the data?

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the explanatory variables
- understand relationships between explanatory and outcome variables
How to look at the data?

- inspect raw data
- summary statistics
- visualize
American community survey

2013 ACS:

- 3M respondents, 87 economic/demographic survey questions
  - income
  - cost of utilities (water, gas, electric)
  - weeks worked per year
  - hours worked per week
  - home ownership
  - looking for work
  - use foodstamps
  - education level
  - state of residence
  - ...

- 1/3 of responses missing

find it at https://people.orie.cornell.edu/mru8/orie4741/data/acs_2013.csv
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited
don’t load a giant file into memory.
your computer will crash.
How do computers work?

on a laptop:

- hard disk: usually ≤ 500 GB
- memory (RAM): usually ≤ 16 GB
- many programs (e.g., Excel): substantially more limited
don’t load a giant file into memory.
your computer will crash.

how big is ACS data?
3M respondents × 100 questions = 300M numbers ≈ 300MB
Inspect raw data

solution for large files: technology from the 70s!

bash shell:

▶ ls -lh
▶ head, tail, less
▶ wc -l
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHTYPE</td>
<td>household type</td>
<td>categorical</td>
</tr>
<tr>
<td>STATEICP</td>
<td>state</td>
<td>categorical</td>
</tr>
<tr>
<td>OWNERSHP</td>
<td>own home</td>
<td>Boolean</td>
</tr>
<tr>
<td>COMMUSE</td>
<td>commercial use</td>
<td>Boolean</td>
</tr>
<tr>
<td>ACREHOUS</td>
<td>house on $\geq$ 10 acres</td>
<td>Boolean</td>
</tr>
<tr>
<td>HHINCOME</td>
<td>household income</td>
<td>real</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>monthly electricity bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>monthly water bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>monthly gas bill</td>
<td>real</td>
</tr>
<tr>
<td>FOODSTMP</td>
<td>on food stamps</td>
<td>Boolean</td>
</tr>
<tr>
<td>HCOVANY</td>
<td>have health insurance</td>
<td>Boolean</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>currently in school</td>
<td>Boolean</td>
</tr>
<tr>
<td>EDUC</td>
<td>highest level of education</td>
<td>ordinal</td>
</tr>
<tr>
<td>GRADEATT</td>
<td>highest grade level attained</td>
<td>ordinal</td>
</tr>
<tr>
<td>EMPSTAT</td>
<td>employment status</td>
<td>categorical</td>
</tr>
<tr>
<td>LABFORCE</td>
<td>in labor force</td>
<td>Boolean</td>
</tr>
<tr>
<td>CLASSWKR</td>
<td>class of worker</td>
<td>Boolean</td>
</tr>
<tr>
<td>WKSWORK2</td>
<td>weeks worked per year</td>
<td>ordinal</td>
</tr>
<tr>
<td>UHRSWORK</td>
<td>usual hours worked per week</td>
<td>real</td>
</tr>
<tr>
<td>LOOKING</td>
<td>looking for work</td>
<td>Boolean</td>
</tr>
<tr>
<td>MIGRATE1</td>
<td>migration status</td>
<td>categorical</td>
</tr>
</tbody>
</table>
Julia and Jupyter

- Julia is a programming language: it parses human-readable code to machine-readable code, executes it, returns the answer.
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as `.ipynb` files.
- Jupyter notebooks display inputs and outputs in a browser.
- JuliaBox is an interface to a webserver running Julia.
Summary statistics

univariate

► mean, median, mode
► max, min, range
► variance
► ... 

explore via Julia + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb
Summary statistics

univariate

▶ mean, median, mode
▶ max, min, range
▶ variance
▶ ...

explore via Julia + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb

multi- (but usually just bi-)variate

▶ correlation, covariance
▶ ...

### The perils of summary statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>
The perils of summary statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
<td>y</td>
<td></td>
<td>y</td>
<td></td>
<td>y</td>
<td></td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
<td></td>
</tr>
</tbody>
</table>

same mean, variance, correlation, line of best fit...
The perils of summary statistics
The perils of summary statistics: modern update

https://www.autodeskresearch.com/publications/samestats
What to visualize?

- examples across all features (usually not)
- plot features across all examples (much more common)
Take away

- always look at (some of) your data
- decide what question you want to answer
Logistics

- OH start today at 3pm. Locations / times are posted on course website.
- Sections are optional, attend any one you prefer. This coming W and the following M and W are all repeats. First one (tomorrow) is a Julia tutorial.
- Canvas is open for submission of hw0, due Thursday 9-5-19.
- Remember to submit participation post on Piazza after each class!
Questions?